in *The Oxford Handbook on Free Will*, ed. R. Kane (Oxford: Oxford University Press), pp 441-60

The Bounds of Freedom

1 Introduction

Are human beings ever really—without qualification—responsible for their actions? Are they ever really *morally* (and not just causally) responsible for their actions? Are they ever *ultimately* responsible for their actions? Are they ever *ultimately morally* responsible for them? Are they ever responsible for their actions in such a way that they are, without any sort of qualification, morally deserving of praise or blame or punishment or reward for them?

This question, with its various strengths, is the only really troublesome question when it comes to the problem of free will, and it is the only question I will consider here. The difficulty with it is simple and well known: there appear to be powerful reasons for answering Yes and powerful reasons for answering No. One might say that there are frames in which the answer is Yes and frames in which the answer is No. I want to draw attention to the fundamental frame in which the answer is No. The point I have to make is old and simple and a priori and I will articulate it in more than one way, as a kind of exercise.

There are also powerful a posteriori reasons for answering No. No seems unavoidable if Einstein's theory of special relativity is anything like correct, for example—a fact little discussed in recent debate about free will. Einstein reckoned that 'a Being endowed with higher insight and more perfect intelligence, watching man and his doings, would smile about man's illusion that he was acting according to his own free will'.[1] Here, however, I will stick to the a priori point.

Being a priori, it holds good whether determinism is true or false: the issue of determinism is completely irrelevant to the present discussion.[2] For the record, though, determinism is the view that the history of the universe is fixed in such a way that everything that happens is necessitated to happen by what has already gone before in such a way that nothing can happen otherwise than it does. It can also be expressed, more simply, as the view that every event has a cause.[3]

2 Some symbols

In speaking of actions I will restrict attention to fully intentional and consciously deliberated actions (as opposed to reflex actions, say, or habitual or otherwise undeliberated actions); not because these are the only ones for which we judge people to be morally responsible, but

[1] Einstein 1931. For an excellent presentation of the a posteriori point see Putnam 1967 and especially Lockwood 2007, who effectively rebuts Putnam's critics. (When I cite a work I give the original publication date when I can, while the page reference is to the edition listed in the bibliography.)

[2] Actually, it is also completely irrelevant on the terms of the a posteriori argument just mentioned: the generality of the argument from special relativity is such that it makes no difference whether determinism is true or false.

[3] Some think that this simple formulation won't do; they think it better to say that determinism is the view that every event *and every aspect of every event* has a cause. But this adjustment is unnecessary, because anything that is characterized as an aspect of an event given one way of individuating events can itself be characterized as an event given another equally good way of individuating events.

because any successful case for the view that people can (without qualification) be morally responsible for their actions must cover these cases, and the other cases raise no fundamentally different questions. I will use 'R' to abbreviate 'truly and without qualification responsible' and the corresponding noun, 'D' to abbreviate 'truly and without qualification deserving of praise or blame or punishment or reward' and the corresponding noun, 'U' to abbreviate 'ultimate' when prefixed to a noun and 'ultimately' when prefixed to an adjective, 'M' to abbreviate 'moral' and 'morally', and '[φ → ψ]' to represent 'φ entails ψ'. I will take it that R and D can be fused to form a single notion—true, unqualified responsibility-and-deservingness or 'RD', for short —in the present context of debate, and I will also use 'RD' as an adjective, meaning '(truly and without qualification) responsible and deserving of praise or blame or punishment or reward'.[4]

With these provisions, the opening question is

Are human beings ever RD for their actions? Are they ever *U*RD for their actions? Are they ever U*M*RD for their actions?

But one of these letters is not needed. With one exception, I will in what follows consider only questions of *moral* responsibility and deservingness, so 'M' can be dropped and taken as read.[5] The question, then, is:

Are human beings ever really RD? Are they ever really URD?[6]

This question raises several others: What exactly is URD? Is there really any interesting distinction to be drawn between RD and URD? Given that [URD → RD], is it also true that [RD → URD]? I will consider these questions in §§4 and 5. Until then I will rely on the reader's pre-reflective understanding of RD and URD and give four versions of the argument—the *Basic Argument*—for answering No to the key question: Are we ever really RD, or URD?

3 The Basic Argument

The Basic Argument has various expressions, but its core is simple and can be quickly stated.

Version 1

1.1 When you act, you do what you do—in the situation in which you find yourself[7]—because of the way you are.

[4] Some actions are neutral in such a way that their performers are not D even if they are R. The idea behind the single notion of RD is that if one is RD then if one is R for some action A then one is also and ipso facto D for A *if* any praise or blame attaches to actions of A-type actions—which it may not do.
[5] I will also regularly omit the phrase 'for their actions'.
[6] The notion of URD is effectively the same as Kane's notion of UR; see THIS BOOK 000.
[7] I will take this qualification for granted.

1.2 If you do what you do because of the way you are, then in order to be URD for what you do you must be URD for the way you are.

But

1.3 You cannot be URD for the way you are.

So

1.4 You cannot be URD for what you do.

Version 1 of the Basic Argument has three premisses, **1.1**, **1.2**, and **1.3**. I take premiss **1.1** to be obvious and will not defend it. I think that **1.2** and **1.3** are also obvious, but I will give them—or close cousins of them—some explicit defence below.

The Basic Argument can be restated as follows.

Version 2

2.1 One cannot be *causa sui*—one cannot be the cause of oneself.

But

2.2 One would have to be *causa sui*, at least in certain crucial mental respects, in order to be URD for one's thoughts and actions.

It follows that

2.3 One cannot be URD for one's thoughts or actions: one cannot be ultimately morally deserving of praise or blame for one's thoughts or actions or one's character or indeed for anything else.

But

2.4 [RD → URD]; *unqualified* responsibility and deservingness requires *ultimate* responsibility and deservingness.

So

2.5 One cannot be RD: one cannot be (truly and without qualification) morally deserving of praise or blame: not for one's thoughts, or actions, or character, or anything else.

This argument goes through whether determinism is true or false, for we cannot be URD either way. Nor, therefore, can we be RD. Even if the property of being *causa sui* is allowed to belong (entirely unintelligibly) to God, it cannot be plausibly supposed to be possessed by ordinary human beings: '*No one* is accountable for existing at all, or for being constituted as

he is, or for living in the circumstances and surroundings in which he lives', as Nietzsche remarked:[8]

> the *causa sui* is the best self-contradiction that has been conceived so far; it is a sort of rape and perversion of logic. But the extravagant pride of man has managed to entangle itself profoundly and frightfully with just this nonsense. The desire for 'freedom of the will' in the superlative metaphysical sense, which still holds sway, unfortunately, in the minds of the half-educated—the desire to bear the entire and ultimate responsibility for one's actions oneself, and to absolve God, the world, ancestors, chance, and society—involves nothing less than to be precisely this *causa sui* and, with more than Baron Münchhausen's audacity, to pull oneself up into existence by the hair, out of the swamps of nothingness.[9]

Version 2 of the Basic Argument has three premisses, **2.1**, **2.2**, and **2.4**. Few would dispute **2.1**, but **2.2** and especially **2.4** can be challenged. I will consider these challenges after setting out a third, longer version of the Basic Argument.[10]

Consider a particular action or piece of deliberation that you engage in, and consider everything about the way you are when you engage in it that leads you engage in it in the way you do. I will call the particular action or piece of deliberation that you engage in '**A**', and I will call everything about the way you are mentally when you engage in it that leads you engage in it in the way you do '**N**'. I will use URD**A**(t) and URD**N**(t) to mean URD for **A** at time t and URD for **N** at time t respectively.

Version 3

3.1 When you act or deliberate, at t_1—when **A** occurs, at t_1—you do what you do, in the situation in which you find yourself, because of the way you are—because you are **N**, at t_1.

This is the first premiss of the argument. I take it to be incontrovertible, quibbles aside, and will not defend it (remember that **N** covers all aspects of the way you are at t; there is no special stress on whatever part of **N** might be identified as your character or personality).

It appears to follow immediately that

3.2 If you are to be URD**A**(t_1)—URD for what you *do*, at t_1, then you must be URD**N**(t_1)—URD for the way you *are*, at t_1 at least in certain crucial mental respects.

(Comment: I take the qualification 'at least in certain mental respects' for granted from now on. Obviously you don't have to be responsible for the way you are in all respects. You don't have to be responsible for your height, age, sex, and so on. But it does seem that you have to be responsible for the way you are mentally at least in certain respects. After all, it is your overall mental make up that leads you to do what you do when you act or deliberate.)

---

[8] Nietzsche 1888: §6.8 ('The Four Great Errors'). For an outstanding discussion of Nietzsche's views on fate and the possibility of self-creation see Leiter 1998.
[9] 1886: §21.
[10] For variants see e.g. G. Strawson 1986: 28-30, 1994a: 6-7, 12-14, 1998: 746-7.

The move from **3.1** to **3.2** can be set out as an explicit premiss:

3.3 [**3.1** → **3.2**]: if, when **A** occurs, you do what you do because you are **N**, because of the way you are, then if you are to be URD**A**(t_1) you must somehow be URD**N**(t_1).[11]

(Comment: **3.3** has deep intuitive plausibility, and I will take it for granted for the moment. Note that **3.2** follows from **3.1** and **3.3**, so that we only have two premisses so far.)

But

3.4 You can't be URD**N**(t_1)—you can't be URD for the way you are in any respect at all, or at any time.

So

3.5 You certainly can't be URD**A**(t_1)—for what you do, at t_1.

This completes the first stage of Version 3. It has three premisses, **3.1**, **3.3**, and **3.4**. I take **3.1** to be incontrovertible, like **1.1** and **2.1**. The second stage of Version 3 is devoted to establishing **3.4**. **3.3** is reserved for discussion in §4.

So far, perhaps, so good. But why is **3.4** true? Why can't you be URD**N**, at least in certain mental respects? Well,

3.6 If it is true that you are URD**N**(t_1)—URD for the way you are, at t_1, in certain mental respects, then it must be true that you have somehow intentionally brought it about that you are **N** at some time t_0 prior to t_1.

(Comment: **3.6**. is another premiss I will not defend, on the grounds that it is evident on reflection. It does not just state that you must have caused yourself to be the way you are, mentally speaking, at least in certain mental respects; that is certainly not enough for ultimate responsibility. It states that you must have consciously and explicitly decided on a way to be and—roughly—must have acted on that decision with success.)

Is it possible for you to have intentionally brought it about that you are **N** at some time t_0 prior to t_1, as **3.6** requires? Well, let us assume that it is. Let us simply assume, for the sake of argument, that

3.7* You have somehow intentionally brought it about that you are **N** at t_0 prior to t_1.

Or rather, more richly, let us simply assume that

3.7 You have somehow intentionally brought it about that you are **N**, at t_0, in such a way that you can now be said to be URD for being **N**, at t_1

[11] This corresponds to **1.2**. I am grateful to Karin Boxer for demanding that I make it explicit.

without enquiring into how exactly this might have come about. [12] Clearly, for **3.7** to be true

3.8 You must already have had a certain mental nature—call it **M**—at t_0, in the light of which you intentionally brought it about that you now have nature **N**.

Why? Because

3.9 If you didn't already have a certain mental nature, at t_0, then you can't then have had any intentions or preferences at all; and if you didn't then have any intentions or preferences at all, you can't be held to be RD, let alone URD, for intentionally bringing anything about, at t_0.

(Comment: I take this premiss too to be evident.)

So **3.8** is true. But there is more to say, because

3.10 For it to be true that you and you alone are RD**N** or URD**N**, at t_1, you must have been RD**M** or URD**M** at t_0—RD or URD for your having had that nature **M** in the light of which you intentionally brought it about that you now have **N**.

(Comment: I take it that this follows from **3.3**, and leave aside the difficulties about the nature of time raised by the work cited in note 1.)

But

3.11 For you to have been RD**M** or URD**M** you must have intentionally brought it about that you had **M**.

(Comment: This is a version of **3.6**.)

So

3.12 You must have intentionally brought it about that you had **M**.

But in that case

3.13 You must (given **3.9**) have existed already with a prior nature, **L**, in the light of which you intentionally brought it about that you had that nature, **M**, in the light of which you intentionally brought it about that you now have nature **N**.

3.14 And so on.

[12] The limiting case of this, presumably, would be the case in which you simply endorsed your existing mental nature **N** from a position of power to change it.

Here one is setting off on a potentially infinite regress: it seems, quite generally, that if one is to be URD**N**, URD for *how one is*, in such a way that one can be URD**A**, URD for *what one does*, something impossible has to be true. There has to be, but there cannot be, a starting point in the series of acts or processes of bringing it about that one is a certain way, or has a certain nature, a starting point that constitutes an act or process of ultimate self-origination. It follows that **3.7** is impossible; in which case **3.4** is true, given **3.6**.

This completes the second stage of Version 3 of the Basic Argument. It assumes **3.7** for reductio and has two premisses, **3.6** and **3.9**, both of which seem evident. As a whole, Version 3 sets out in more detail the claim of Versions 1 and 2—the claim that URD requires the occurrence of processes of ultimate self-origination of a kind that are impossible. Hardly any of those who appear to believe in URD—nearly all human beings[13]—have ever had any conscious thought to the effect that it requires some such ultimate self-origination, but that is beside the point.[14]

In §4, the next section, I will look at two of the premisses (or premiss-groups) of the various versions of the Basic Argument. In §5 I will say something more about what URD is meant to be. In §6 I will consider a different challenge to one of the premiss-groups of the Basic Argument. I will end this section with a more everyday version of the Basic Argument.

Version 4

4.1 Initially—early in life—one is the way one is as a result of one's heredity and experience.[15]

4.2 One's heredity and early experience are *obviously* things for which one cannot be held to be in any way RD or URD.[16]

4.3 One cannot at any later stage of one's life hope to accede to URD for the way one is, and, in particular, for the way one is morally speaking, by trying to change the way one already is as a result of one's heredity and previous experience.

4.4 There is no other way in which one could hope to accede to URD for how one is.

So

4.5 One cannot be URD for how one is in any way at all.

And if

---

[13] For a recent exposition of this point see Smilansky 2000; and **VII** below.
[14] Some of course do have the conscious thought. See **VI** below.
[15] I take 'experience' to include all impacts or effects on the mind, where this includes internal bodily impacts as well as external environmental impacts.
[16] This might not be true if there were reincarnation, but reincarnation would just shift the problem backwards—we would be off on another regress.

4.6 [RD → URD]

as supposed on page 3 (**2.4**), then

4.7 One cannot be RD how one is in any way at all.

I take **4.1** and **4.2** to be evident, assume **4.4**, and discuss **4.6/2.4** in the next section. I will now defend **4.3**.

4.8 The reason **4.3** is true is not that one cannot try to change the way one is as a result of one's heredity and previous experience, or that one cannot succeed if one does try. One can both try and succeed. The reason **4.3** is true is simply that if one does try to change oneself then one aims at the particular changes one does aim at, and takes the particular steps one does take in the attempt to bring them about, and succeeds in bringing them about to the extent that one does, in the situations in which one finds oneself, wholly because of the way one already is as a result of one's heredity and previous experience—which is something for which one is in no way URD.

4.9 There may be certain further changes that one can bring about only after one has brought about certain initial changes, and one may succeed in bringing about some of these further changes too. But the point made in **4.8** simply reapplies.

Note that once again it makes no difference whether determinism is true or false. If determinism is false, it may be that some changes in the way one is are traceable to the influence of indeterministic or random factors. It is even possible that difficult decisions or efforts to change oneself may trigger indeterministic goings on in the brain.[17] But indeterministic or random factors, for whose particular character one is *ex hypothesi* in no way responsible, cannot contribute in any way to one's being URD for how one is.[18]

The claim, then, is not that people cannot change the way they are. They can, in certain respects. It is only that people cannot be supposed to change themselves in such a way as to be or become URD for the way they are, and hence for their actions. One can put the point by saying that the way you are is, ultimately, in every last detail, a matter of luck—good or bad.[19]

'Character is fate: your character determines your fate. So radical freedom is excluded', say Heracleitus, Novalis, George Eliot and others. 'Not so fast', say the Sartreans: 'Character may determine fate, but character is choice. Character is a product of choice, so you can choose your fate. Radical freedom is possible after all.' 'Maybe character is choice', reply proponents of the Basic Argument, taking the side of Heracleitus, 'but choice is character: character determines choice, even choice of character. And character is fate. So radical freedom is excluded after all.'

---

[17] See Kane 1996: SUPPLY REF.

[18] Compare Kane 1989, 1996. I state my differences with Kane in Strawson 1994a: 17-21 and Strawson 2000: 149-155.

[19] There is a sense in which talk of luck is odd in this context (see e.g. Hurley 2002), but it makes the point clearly.

4 Two premisses, three positions

So much for the Basic Argument. I want now to consider two of its premisses. First, **2.4/4.6**. Is it true, or even plausible, that [RD → URD]?

Some say No. Faced with arguments like those just given, they take the following position:

Position 1. There is indeed an ineliminable sense in which human beings cannot be URD**N**, and it does indeed follow that there is an ineliminable sense in which they cannot be URD**A**. But who needs the 'U', the 'ultimate'? Even if these sorts of ultimacy are unavailable, human beings can be truly RD**A** in such a way as to be wholly proper objects of moral praise and blame and punishment and reward.

One popular version of this position runs as follows:

Position 2. Being RD**A**, *fully*, *wholly* and without qualification responsible for some action **A**, is just a matter of being a responsible (as we naturally say) adult, a fully responsible adult: a normal self-conscious adult human being who is not subject to any compulsion, so far as **A** is concerned. That is all. Being a normal self-conscious adult human being is already sufficient for RD**A**, whatever else is or is not necessary for it, and since we know such adult human beings exist we also know that RD**A** is possible and actual. No metaphysical issues need be considered. Philosophers can distinguish URD**A** from RD**A** if they like. They can raise complicated questions about whether one can be URD**N** or RD**N**—URD or even merely RD for how one is. Let them. RD**A** is possible and actual whatever scintilla-loving philosophers choose to say about RD**N** and URD**N**. Have they defined URD**A** in such a way that it (URD**A**) is neither actual nor possible? Let them. RD**A** is possible and actual for all that.

Some go further, and reject the possibility of any gap between RD**A** and URD**A**.

Position 3. Look, anything that really counts as *genuine* or *full* or unqualified RD**A** just is URD**A**. The adjective 'U' or 'ultimate' adds nothing. RD**A** certainly exists, and RD**A** is RD**A** is URD**A**. Suppose there is a clear and undeniable sense in which human beings cannot be URD**N**; it just doesn't follow that they can't be URD**A**. RD**A** exists and RD**A** = URD**A**. The idea that there might be some further kind of radical, 'ultimate' responsibility for action over and above the kind of straight-up responsibility possessed by a normal self-conscious adult human being is moonshine.

I disagree. It is possible to characterize a notion of URD**A** that is importantly distinct from any notion of RD**A** truly applicable to human action; I will do so in the next section. And yet I agree that there is a very important way of understanding the notion of RD given which it is true that human beings can—rightly and without reservation—be held to be fully RD for their actions. And I agree that this notion of RD allows us to say that human beings can be fully RD for their actions even if they are not *U*RD either for how they are or for their actions. This notion of RD is a *compatibilist* notion. Compatibilists have laid out its structure

and variants with great ingenuity and devotion over many years,[20] and I have nothing to add to what has been said about it. My present task is simply to provide a reminder of what compatibilism is not and cannot be, in case anyone should have any tendency to forget: a reminder of the fact that compatibilism is nothing more than a 'wretched subterfuge. . . , a petty word-jugglery',[21] 'so much gobbledegook',[22] when it is taken to be more than it is.[23]

I will return to the question whether [RD → URD] in various ways. First, I want to mention the second premiss (the **1.2**, **2.2**, **3.2-3.3** group), which can be expressed as [URD**A** → URD**N**]. I have endorsed it, argued that URD**N** is impossible, and concluded that URD**A** is impossible. I don't really think that it needs defence, but the characterization of URD in the next section can be taken as a defence if one is felt to be needed.

Robert Kane endorses the [URD**A** → URD**N**] premiss, but he argues that there is a sufficient sense in which URD**N** is possible, and that there is (therefore)[24] also a sufficient sense in which URD**A** is possible. He further holds that URD**N** is possible only if determinism is false, adopting an explicitly incompatibilist—libertarian—position.

Immanuel Kant agrees with Kane and me in accepting that [URD**A** → URD**N**], and he agrees with Kane, but not me, in asserting that URD**A** is possible. He goes further than both of us in asserting that it is knowably actual. Unlike Kane, however, he does not think that one can give any substantive account of how URD**N** is possible.[25]

Other positions are of course possible.[26] Most, though, are likely to protest that questions about whether or not we are or can be responsible for how we are are simply (even magnificently) irrelevant to any and all assessments of RD that actually concern us.[27]

Could this be true? It is certainly true that such questions seem irrelevant in most ordinary moral discussions, and if they are irrelevant then the whole issue of whether or not the [URD**A** → URD**N**] premiss is true is equally irrelevant.

I will reject the charge of irrelevance in §6, and make three suggestions about what motivates it in §7. First, though, I must say something more about what I take URD to be.

5 Ultimate responsibility

[20] Cf. e.g. Hobbes 1651, Locke 1690, Hume 1748, Schlick 1930, Hobart 1934, Frankfurt 1971, Watson 1975, Fischer 1994—and many others.

[21] Kant 1788: 191 (Ak. V. 97).

[22] Anscombe 1971: 146.

[23] On this issue, see Smilansky 2000, especially Part 2.

[24] This 'therefore' also requires [URD**N** → URD**A**], the converse of the premiss that Kane and I agree on. But [URD**N** → URD**A**] is clearly very plausible. Nagel notes its plausibility explicitly (1987: 00) when commenting on a doubt that I raise about it in *Freedom and Belief* (1986: 299-301; I propose [1] that one must have a positive *sense* of oneself as URD**A** in order to be URD**A** *sans phrase*, [2] that one might conceivably lack any such sense of oneself as URD**A** even if URD**N** were possible and even if one were in fact URD**N**, concluding [3] that [URD**N** → URD**A**] is to that extent not true).

[25] In various places he claims that we can know that URD**A** is actual even though we cannot even comprehend its possibility, and he would presumably take exactly the same line about URD**N**. Cf. e.g. Kant 1785: 127 (Ak. IV. 459), 1788: 4 (Ak. V. 4), 1793: 45 n (Ak. VI. 49-50).

[26] Some, perhaps, may concede that we cannot be *U*RD**N** while insisting that we can none the less be RD**N** in some robust way—so that we can be RD**A** even if [RD**A** → RD**N**].

[27] Even those who reject all forms of compatibilism may take this view. C. A. Campbell (1967), for example, is a libertarian who takes it that URD**A** is possible even if URD**N** is not.

What exactly is URD, this 'ultimate' responsibility that is meant to be impossible? One simple and dramatic way to characterize it is by reference to the story of heaven and hell. URD is *heaven-and-hell* responsibility: if we have URD then it makes sense to propose that it could be just—without any qualification—to punish some of us with (possibly everlasting) torment in hell and reward others with (possibly everlasting) bliss in heaven. The proposal is morally repugnant, but it is perfectly intelligible because if we really have URD then what we do is wholly and entirely up to us in some absolute, buck-stopping way.

One does not have to believe in the story of heaven and hell in order to understand the notion of URD it is used to illustrate. Nor does one have to believe in the story of heaven and hell in order to believe in URD (many atheists have believed in URD). One doesn't even have to have heard of the story, which is useful here simply because it illustrates the *kind* of absolute or ultimate responsibility—URD—that many suppose themselves to have. And the core notion of URD has no essential connection with moral matters. If we temporarily drop the 'M' for 'moral' that is implicit in 'URD' (p. 000) we may observe, first, that self-conscious agents that face difficult life-determining choices but that have no conception of morality at all can have a sense of UR—of radical, absolute, buck-stopping (buck-printing) 'up-to-me-ness' in choice and action—that is just as powerful as ours, and, second, that the story of heaven and hell can be used to convey the absolute character of this non-moral URD just as well as it conveys the absolute character of any moral URD.

So much for the notion of URD. There is a sense in which it is not coherent, but it does not follow that it is unintelligible or has no genuine content. That could not be, for it is a notion that is central to common moral consciousness, at least in the West, and certainly not just in the West. I have conveyed its content by reference to the story of heaven and hell, but it can also be conveyed less colourfully as follows: URD is responsibility and desert of such a kind that it can exist if and only if punishment and reward can be fair or just without having any pragmatic justification, or indeed any justification that appeals to the notion of distributive justice.[28]

Whichever characterization one prefers, it is precisely (only) because one has a grasp of the content of the notion of URD that one can see, or can be brought to see, that it is incoherent. It is the same with the notion of a round square. Some may say that they don't really know what the content of this notion is, but it is easy to specify. A round square is an equiangular, equilateral, rectilinear, quadrilateral closed plane figure every point on the periphery of which is equidistant from a single point within its periphery. It is because we know the content of the notion that we know that there cannot be such a thing as a round square, and the same is true of the notion of URD. Many say that statements or concepts that are self-contradictory are meaningless, but meaningfulness is a necessary condition of contradictoriness.

[28] The qualification referring to distributive justice is strictly speaking unnecessary. Suppose X's deliberate and intentional action gives rise to a collective burden that can be alleviated only by imposing a special burden on some member of the community; or suppose the performance of the action has the consequence that someone must bear a burden whether anyone likes it or not. And suppose X knows that this will be so. Then even if it is thought to be intrinsically fair or just—in some absolute, wholly unqualified sense—to impose the burden on X, it doesn't follow that there is any way in which the burden can correctly be thought of as a fair or just *punishment*.

—'You aren't making any progress in offering these characterizations because both of them make use of some notion of 'ultimate' justice, and exactly the same sort of common-sense move that was made in response to the qualification of 'responsibility' by 'ultimate' can be made in the case of the qualification of 'justice' by 'ultimate'. Human beings cannot be URDA given your characterization of URD, but praise and punishment of, and reward and blame for, human action can none the less be just, just *tout court*, just without any qualification. Other things being equal, to be capable of being justly punished, justly punished *sans phrase*, is just a matter of being a normal self-conscious adult human being who is not subject to any relevant compulsion. Your attempt to characterize URD in terms of justice just doesn't work.'

This objection simply restates positions 1 and 2 (p. 9) in terms of justice instead of RD. It allows the sense in which we are not URD**A**, but claims that punishment on moral grounds can none the less be just *sans phrase*, heaven-and-hell just, just without any qualification or appeal to pragmatic considerations or considerations of distributive justice. I disagree. We may have reached the end of argument.

6 The Relevance View

—'References to RD**N** and URD**N**—I will use '/RD**N**/' to refer to them jointly when the distinction between them is not at issue—disappeared from the discussion in the last section. References to URD**A** and RD**A**—'/RD**A**/' for short—did not. Doesn't this strongly confirm the view that questions about /RD**N**/ are irrelevant to any of the issues about /RD**A**/ that actually concern us in everyday life? And aren't such questions equally irrelevant to sensible moral philosophy? And aren't they irrelevant to sensible moral philosophy precisely because they're irrelevant to the issues about /RD**A**/ that actually concern us in everyday life?'

No, in answer to all these questions.

—'But even if questions about /RD**N**/ aren't irrelevant to the issues about /RD**A**/ that concern us in everyday life, they're generally thought to be irrelevant. The *Irrelevance View*, as one might call it, lies deep in ordinary moral thought and feeling.'

This is an important fact, and I will try to explain it in §7. But it is equally important that the directly contrary view—the *Relevance View*, according to which /RD**A**/ does somehow involve /RD**N**/, so that questions about /RD**N**/ are profoundly relevant to issues of moral responsibility—also lies deep in ordinary moral thought and feeling, constantly ready to precipitate out into consciousness in ways which I will consider now.[29]

One way in which the Relevance View manifests itself is in the sense that many have that they are somehow or other responsible for, answerable for, how they are mentally, or at least for certain crucial aspects of how they are mentally. Certainly we do not ordinarily suppose that we have actually gone through some sort of active process of self-determination at some particular past time. And yet it seems accurate to say that we do unreflectively experience

[29] It takes the distortions that arise from a philosophical training to doubt this obvious fact. It also takes a philosophical training to be confused enough—as some compatibilists have been—to suppose that it takes a philosophical training to think that this fact is obvious.

ourselves, in many respects, very much as we would experience ourselves if we did think we had engaged in some such process of self-determination, or had at least engaged in some process of scanning and ratification of how we are mentally that we had undertaken from a position of power to induce change. Many, perhaps, feel that it is just a fact about growing up that one comes to be such that one is /RD**N**/.

Some find traits in themselves that they regret, or experience as foreign, and feel powerless to change. This, however, does not put the present point in doubt, for traits can appear as regrettable or foreign only against a background of character traits that are not regretted or experienced as foreign, but are, rather, identified with. In general, people have a strong sense of general identification with their character (it may well be strengthened, not weakened, by the experience of some tendency as alien), and this identification seems to carry within itself a powerful implicit sense that one is, generally, somehow in control of, and in any case answerable for, how one is.[30]

So /RD**N**/ does not always appear irrelevant in ordinary moral thought. And the idea that /RD**N**/ is necessary for /RD**A**/ arises with intense naturalness, and in an explicit form, when people begin to reflect about the nature of moral responsibility, as they quite often do. Many who feel certain that they are URD**A** also explicitly hold that [URD**A** → URD**N**], and are accordingly sure that they are URD**N**.[31] John Patten, British Minister for Education in the 1980s, a non-philosopher and a Roman Catholic, thinks it 'self-evident that as we grow up each individual chooses whether to be good or bad'. E. H. Carr, a historian, holds that 'normal adult human beings are ultimately responsible for their own personality'. Among professional philosophers Jean-Paul Sartre speaks of 'the choice that each man makes of his personality', and holds that 'man is responsible for what he is', and Robert Kane is explicit about the point that one must show that URD**N** is possible in order to show that URD**A** is possible. Immanuel Kant puts the view very clearly when he claims that

> man *himself* must make or have made himself into whatever, in a moral sense, whether good or evil, he is to become. Either condition must be an effect of his free choice; for otherwise he could not be held responsible for it and could therefore be *morally* neither good nor evil,

and since he is committed to belief in URD**A**, he takes it that such self-creation does indeed take place, writing accordingly of 'man's character, which he himself creates' and of the 'knowledge [that one has] of oneself as a person who . . . is his own originator'. Aristotle also seems to take this view for granted.[32]

7 The Irrelevance View and the Agent-Self

[30] It is hardly surprising that there is some such sense of identification, because the subjects who contemplate their own character sets are actually constituted, character-wise and pro-attitude-wise, by the very character sets they are considering. See Strawson 1986: 111-113.

[31] One common progress of thought is from [1] an unquestioned conviction that people have URD**A** to [2] the thought, after a little reflection, that URD**A** requires URD**N**, to [3], the conviction—whose examination is shied away from—that URD**N** is possible, actual, and standard.

[32] Carr 1961: 89; Sartre 1948: 29, and in the *New Left Review* 1969 (quoted in Wiggins, 1975); Kant 1793: 40 (Ak. VI. 44), 1788: 101 (Ak. V. 98); Patten in *The Spectator*, January 1992; Aristotle, *Nicomachean Ethics* V.3. Among recent discussions see e.g. Anglin 1990, Gomberg 1975, Honderich 1993, Klein 1990, Pereboom 1995, Smilansky 2000, Sorabji 1980 (on Aristotle).

So much for the Relevance View. How does the contrary view (that URD**N** is irrelevant to URD**A)** manifest itself? The primary fact is this: it seems that we naturally take it that our capacity for fully explicit self-conscious deliberation in a situation of choice—our capacity to be to be explicitly aware of ourselves as facing choices and engaging in processes of reasoning about what to do—suffices by itself to constitute us as /RD**A**/ in the strongest possible sense. Should the issue of /RD**N**/ be raised—and it standardly isn't—one is likely to feel that one's full self-conscious awareness of oneself and one's situation when one chooses simply vapourizes any supposed consequences of the fact that one neither is nor can be URD**N**. It seems as if the mere fact of one's self-conscious presence in the situation of choice confers radical, total /RD**A**/ on one—it seems obvious that it does so. One may in the final analysis be wholly constituted as the sort of person one is by factors for which one is not and cannot be in any way URD, and one may acknowledge this, but the threat that this fact is alleged to pose to one's claim to /RD**A**/ seems to be annihilated by the simple fact of one's full self-conscious awareness of one's situation.[33]

I think this correctly describes one of the forms taken by our powerful belief in URD**A**. It is not, however, an account of anything that could really constitute URD**A**, for reasons already given: when one acts after explicit self-conscious deliberation, one acts for certain reasons. Which reasons finally weigh with one is wholly a matter of one's mental nature **N**, which is something for which one cannot be in any way URD.

The conviction that fully explicit self-conscious awareness of one's situation can nonetheless be a sufficient foundation of URD**A** is *extremely* powerful; it runs deeper than rational argument, and seems to survive untouched, in the everyday conduct of life, even after the validity of the argument against URD**A** has been admitted; but that is no reason to think that it is correct.

Suppose you arrive at a shop on the evening of a national holiday, intending to buy a cake with your last ten pound note to supplement the generous preparations you have already made.[34] Everything is closing down. There is one cake left; it costs ten pounds. On the steps of the shop someone is shaking an Oxfam tin. You stop, and it seems clear to you that it is entirely up to you what you do next—in such a way that you will be RD**A** and indeed URD**A** for whatever you do do. The situation is in fact *utterly* clear: you can put the money in the tin, or go in and buy the cake, or just walk away. You are not only completely free to choose in this situation. You are not free not to choose. You are condemned to freedom, in Sartre's phrase. You are already in a state of full consciousness of what is (morally) at stake and you cannot prescind from that consciousness. You cannot somehow slip out of it. You have to choose. You may be someone who believes that determinism is true: you may believe that in five—two—minutes time you will be able to look back on the situation you are now in and say, of what you will by then have done, 'It was determined that I should do that'. But even if

[33] 'To observe a child of two fully in control of its limbs, doing what it wants to do with them, and to this extent fully free to act in the compatibilist sense of this phrase, and to realize that it is precisely such unremitting experience of self-control that is the deepest foundation of our naturally *in*compatibilistic sense … of URD**A**, is … to understand one of the most important facts about the genesis and power of our ordinary strong [incompatibilistic] sense of freedom' (Strawson 1986:111).

[34] I have told this story before in Strawson 1986: vii, Strawson 1998: §4.

you do fervently believe this, it does not check your current sense of your URD**A** in any way.[35]

One diagnosis of this phenomenon is that one can't really accept or live the rather specific and theoretical thought that determinism may be true, in such situations of choice, and can't help thinking that the falsity of determinism might make URD**A** possible. But this is too complicated: most people don't think about determinism at all, still less think that its falsity might be necessary for URD**A**.[36] In situations like this one's URD**A** seems to stem simply from the fact that one is fully conscious of one's situation, and knows that one can choose, and believes that one action is morally better than the other. This full awareness seems to be immediately enough to confer URD**A**. And yet it cannot really do so, as the Basic Argument shows. For [URD**A** → URD**N**] and URD**N** is provably impossible.

This raises an interesting question: Must *any* cognitively sophisticated, rational, self-conscious agent that faces choices and is fully aware of the fact that it does so experience itself as being URD**A**, simply in virtue of the fact that it is a self-conscious agent (and whether or not it has a conception of moral responsibility)? It seems that we human beings cannot help experiencing ourselves as URD**A**, but perhaps this is a human peculiarity or limitation, not an inescapable feature of any possible self-conscious agent.[37] And perhaps it is not inevitable for human beings. Krishnamurti is categorical that 'you do not choose, you do not decide, when you see things very clearly …. Only the unintelligent mind exercises choice in life'. A spiritually advanced or 'truly intelligent mind simply cannot have choice', because it 'can . . . only choose the path of truth'. 'Only the unintelligent mind has free will'—by which he means experience of radical free will.

A related thought is expressed by Saul Bellow in *Humboldt's Gift*: 'In the next realm, where things are clearer, clarity eats into freedom. We are free on earth'—i.e. we experience ourselves as radically free—'because of cloudiness, because of error, because of marvellous limitation.' And Spinoza extends the point to God. God cannot, he says, 'be said . . . to act from freedom of the will', and if this is so then (being omniscient) he cannot think that he does so.[38]

This is one way in which ordinary thought moves in support of the view that questions about /RD**N**/ are irrelevant to the issue of /RD**A**/. But it is also very tempted by the idea that /RD**A**/ is possible because one's *self*—i.e. *the* self, the *agent-self*, the thing that one most fundamentally is, both morally speaking and in general—is in some crucial way independent of one's general mental nature **N**, one's overall character, personality, motivational structure. What happens when one faces a difficult choice between X, doing one's duty, and Y, following one's non-moral desires? Well, given **N**, one responds in a certain way. One is swayed by reasons for and against both X and Y. One tends towards X or Y, given **N**. But

---

[35] Note that this description of the character of our experience gives further content or colour to the characterization of URD offered in **V**.

[36] It may be added that the feeling of URD**A** seems to be just as inescapable for someone who has been convinced by the Basic Argument against URD**A** given in **III**, which does not depend on determinism in any way: even clearheaded acceptance of the force of the Basic Argument seems to fail to have any impact on one's sense of one's URD**A** as one stands there, wondering what to do.

[37] See, though, Popper 1949, MacKay 1960, for a general argument that no self-conscious agent can truly experience its choices and actions as determined even if determinism is true. See also G. Strawson 1986: ch. 13 and pp. 281-284; Smilansky 2000: Part II.

[38] Krishnamurti 1983: 33, 204; Bellow 1977: 140; Spinoza 0000: 000.

one is as an agent-self independent of **N**, on this picture of things, and one can be /RD**A**/ in a situation like this even if (even though) one cannot be /RD**N**/, because although one's nature **N** certainly *inclines* one to do one thing rather than another it does not thereby *necessitate* one to do one thing rather than the other.[39] As an agent-self (the thought goes) one incorporates a power of free decision that is independent of all the particularities of **N** in such a way that one can after all count as URD**A** even though one is not ultimately responsible for any aspect of **N**.[40]

That, at least, is the story. The agent-self decides in the light of **N** but is not determined by **N** and is therefore free. But the following question arises: *Why* does the agent-self decide as it does? And the general answer is clear. Whatever the agent-self decides, it decides as it does because of the overall way it is; it too must have a nature—call it **N*—**of some sort. And this necessary truth returns us to where we started. Once again it seems that the agent-self must be responsible for **N***—URD**N*** or RD**N***—in order to be URD**A**. But this is impossible, for the reasons given in **3**: nothing can be *causa sui* in the required way. Whatever the nature of the agent-self, it is ultimately a matter of luck (or grace, as some would have it) that it is as it is.

It may be proposed that the agent-self decides as it does partly or wholly because of the presence of indeterministic occurrences in the decision process. But this is no good because it is as clear as ever that indeterministic occurrences can never be a source of URD**A**.[41] The story of the agent-self may add another layer to the description of the human decision process, and it may have a certain phenomenological aptness, considered as such a description, but it cannot change the fact that human beings cannot be /RD**N**/ in such a way as to be /RD**A**/.[42]

It cannot, in other words, change the fact that human beings can never be truly or without qualification morally responsible for their actions, responsible for them in such a way that they are flat-out deserving of moral praise or blame or punishment or reward for them. This is, in a sense, a quite bewildering fact. But it is a fact none the less. We are what we are, and we cannot be thought to have made ourselves *in such a way* that we can be held to be free in our actions *in such a way* that we can be held to be RD for our actions *in such a way* that any punishment or reward for our actions is ultimately just or fair. Punishments and rewards can seem intrinsically appropriate or profoundly fitting to us in spite of this, and many of the various institutions of punishment and reward in human society seem both beneficial and practically indispensable. But if one takes the notion of justice that is central to our intellectual and cultural tradition seriously, the evident consequence of the Basic Argument is that there is a fundamental sense in which no punishment or reward is ever ultimately just. It is *exactly* as just to punish or reward people for their actions as it is to punish or reward them for the (natural) colour of their hair or the (natural) shape of their faces.

---

[39] The distinction is Leibniz's (1686).

[40] C. A. Campbell (1967) gives philosophical expression to this view.

[41] See notes 000 and 000 above.

[42] Another a posteriori argument cuts in at this point: even if some notion of the agent-self is defensible there are powerful neurophysiological reasons for thinking that the 'conscious self' or 'conscious I' cannot be supposed to be the author of decisions and initiator of actions. See Norretranders 1991: ch. 9, and, for the work on which Norretranders draws, Libet 1985, 1987. (See also Libet 1999, a piece which contains considerable conceptual confusion.

There is much more to say about free will, and the point made in this paper is just the beginning. But it is the beginning. It is important to be clear about it, and to try not to avoid or occlude it in any way.

References

Anscombe, G. E. M. 1971. 'Causality and Determination: An Inaugural Lecture.' Cambridge University Press.
Aristotle c 330 BCE/1953. *Nicomachean Ethics*, translated by J. A. K. Thomson. London: Penguin.
Campbell, C. A. 1957. 'Has the Self "Free Will"?' In *On Selfhood and Godhood*. London: Allen and Unwin.
Einstein, A. 1931. 'About free will.' In *The Golden Book of Tagore: A Homage to Rabindranath Tagore from India and the World in Celebration of His Seventieth Birthday*, edited by Ramananda Chatterjee. Calcutta: Golden Book Committee.
Fischer, J. 1994. *The Metaphysics of Free Will: A Study of Control*. Oxford: Blackwell.
Anglin, W. 1990. *Free Will and the Christian Faith*. Oxford: Oxford University Press.
Frankfurt, H. 1988. *The Importance of What We Care About* (essays 1-5). Cambridge University Press.
Gomberg, P. 1975. 'Free Will as Ultimate Responsibility.' *American Philosophical Quarterly* **15**: 205-12.
Hobbes, T. 1651/1996. *Leviathan*, edited by Richard Tuck. Cambridge University Press.
Honderich 1993
Hume, D. 1748/1978. *Enquiry Concerning Human Understanding*. Oxford: Clarendon Press.
Hurley, S. 2002. 'Luck, Responsibility, and the "Natural Lottery".' in *Journal of Political Philosophy* **10**:79–94.
Kane, R. 1989. 'Two Kinds of Incompatibilism.' *Philosophy and Phenomenological Research* **50**: 219-54.
Kane, R. 1996. *The Significance of Free Will* New York: Oxford University Press.
Kant, I. 1793/1960. *Religion within the Limits of Reason Alone*. trans. T. M. Greene and H. H. Hudson New York: Harper and Row.
Kant, I. 1785–6/1948. *Grundlegung zur Metaphysik der Sitten* translated by H. J. Paton as *The Moral Law* (London: Hutchinson, 1948); republished as *Groundwork of the Metaphysic of Morals* (New York: Harper, 1948).
Kant, I. 1785/1993. *Opus posthumum*. Translated by E. Förster & M. Rosen. Cambridge University Press.
Kant, I. 1788. *Kritik der praktischen Vernunft—Critique of Practical Reason* (Riga: Hartknoch).
Kant, I. 1793. *Die Religion innerhalb der Grenzen der blossen Vernunft* (Königsberg: Nicolovius).
Klein, M. 1990
Leibniz, G. 1686/1988. *Discourse on Metaphysics*, translated by R. Martin, D. Niall, and S Brown. Manchester: Manchester University Press.
Leiter, B. 1998. 'The Paradox of Fatalism and Self-Creation in Nietzsche.' In *Willing and Nothingness: Schopenhauer as Nietzsche's Educator*, edited by C. Janaway. Oxford University Press.
Libet, B. 1985. 'Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action.' *Behavioral and Brain Sciences* **8**: 529-566.
Libet, B. 1987. Are the Mental Experiences of Will and Self-control Significant for the Performance of a Voluntary Act?' *Behavioral and Brain Sciences* **10**: 783–786.
Libet, B. 1999. 'Do We Have Free Will?' *Journal of Consciousness Studies* **6**: 47-57.
Lockwood, M. 2007. 'Taking Space-Time Seriously', in M. Lockwood *The Labyrinth of Time: Introducing the Universe*. Oxford: Oxford University Press.
Lutyens, M. 1983. *Krishnamurti: the Years of Fulfilment.* London: John Murray.
MacKay, D. 1960. 'On the Logical Indeterminacy of a Free Choice.' *Mind* **69**: 31–40.
Nagel, T. 1987. 'Is that you, James?' *London Review of Books* **9** (Oct 1).
Nietzsche, F. 1886. *Jenseits von Gut und Böse* (*Beyond Good and Evil*). Leipzig: Naumann.
Nietzsche, F. 1888/2005. *Twilight of the Idols*, ed. A. Ridley and J. Norman, trans. J. Norman. Cambridge University Press.
Norretranders, T. 1991/1998. *The User Illusion: Cutting Consciousness Down To Size*. London: Penguin.
Pereboom, D. 2001. *Living without Free Will*. Cambridge University Press.
Popper, K. 1950. 'Indeterminism in Classical and Quantum Physics.' *Brit. J. Phil. Sci.* **1**: 117-133, 173-195.
Putnam, H. 1967. 'Time and Physical Geometry.' *Journal of Philosophy* 64: 240–247.
Sartre, J.-P. 1948. *Existentialism and Humanism*. London: Methuen.
Schlick, M. 1930/1939. 'When Is a Man Responsible?' In *Problems of Ethics*, trans. D. Rynin. New York: Prentice-Hall.
Smilansky, S. 2000. *Free Will and Illusion*. Oxford: Clarendon Press.
Sorabji, R. 1980. *Necessity, Cause and Blame: Perspectives on Aristotle's Philosophy*. Ithaca: Cornell University Press.
Spinoza, B. 1677/1985. *Ethics*, translated by E. Curley. Princeton University Press.
Strawson, G. 1986/2010. *Freedom and Belief*, 2nd edition. Oxford: Clarendon Press.
Strawson, G. 1994a. 'The Impossibility of Moral Responsibility.' *Philosophical Studies* **75**: 5-24.
Strawson, G. 1994b. *Mental Reality*. Cambridge, MA: MIT Press.

Strawson, G. 1998. ‘Free Will.’ In *The Routledge Encyclopedia of Philosophy*, edited by E. Craig. London: Routledge.
Strawson, G. 2000. ‘The Unhelpfulness of Indeterminism.’ *Philosophy and Phenomenological Research* **60**: 149-156.
Strawson, P. F. 1962. ‘Freedom and Resentment.’ In *Freedom and Resentment*. London: Methuen.
Watson, G. 1975/1982. ‘Free Agency.’ In *Free Will*, edited by G. Watson. Oxford University Press.
Wiggins, D. 1973. ‘Towards A Reasonable Libertarianism.’ In *Essays on Freedom of Action*, edited by T. Honderich. London: Routledge.